

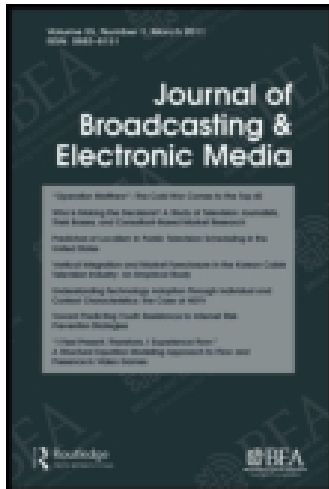
This article was downloaded by: [Chengdu Branch of National Science Library]

On: 12 January 2015, At: 16:59

Publisher: Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954

Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Journal of Broadcasting & Electronic Media

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/hbem20>

### Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods

Seth C. Lewis<sup>a</sup>, Rodrigo Zamith<sup>a</sup> & Alfred Hermida<sup>b</sup>

<sup>a</sup> School of Journalism & Mass Communication at the University of Minnesota-Twin Cities

<sup>b</sup> School of Journalism at the University of British Columbia

Published online: 12 Mar 2013.

To cite this article: Seth C. Lewis, Rodrigo Zamith & Alfred Hermida (2013) Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods, Journal of Broadcasting & Electronic Media, 57:1, 34-52, DOI: [10.1080/08838151.2012.761702](https://doi.org/10.1080/08838151.2012.761702)

To link to this article: <http://dx.doi.org/10.1080/08838151.2012.761702>

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the

Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms & Conditions of access and use can be found at <http://www.tandfonline.com/page/terms-and-conditions>

# Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods

**Seth C. Lewis, Rodrigo Zamith, and Alfred Hermida**

*Massive datasets of communication are challenging traditional, human-driven approaches to content analysis. Computational methods present enticing solutions to these problems but in many cases are insufficient on their own. We argue that an approach blending computational and manual methods throughout the content analysis process may yield more fruitful results, and draw on a case study of news sourcing on Twitter to illustrate this hybrid approach in action. Careful combinations of computational and manual techniques can preserve the strengths of traditional content analysis, with its systematic rigor and contextual sensitivity, while also maximizing the large-scale capacity of Big Data and the algorithmic accuracy of computational methods.*

The term *Big Data* is often invoked to describe the overwhelming volume of information produced by and about human activity, made possible by the growing ubiquity of mobile devices, tracking tools, always-on sensors, and cheap computing storage. “In a digitized world, consumers going about their day—communicating, browsing, buying, sharing, searching—create their own enormous trails of data” (Manyika et al., 2011, p. 1). Technological advances have made it easier than ever to harness, organize, and scrutinize massive repositories of these digital traces; computational techniques for large-scale data analysis that once required supercomputers now can be deployed on a desktop computer (Manovich, 2012). This development has created new opportunities for computational approaches to social

---

**Seth C. Lewis** (Ph.D., University of Texas at Austin) is an assistant professor in the School of Journalism & Mass Communication at the University of Minnesota–Twin Cities. His research explores the social implications of information technology and digital media for the dynamics of media work and media innovation, particularly in the context of journalism.

**Rodrigo Zamith** is a doctoral student in the School of Journalism & Mass Communication at the University of Minnesota–Twin Cities. His research interests include computational research methods in content analysis, journalism and new media technologies, and the interplay between issue framing, media, and policy.

**Alfred Hermida** is an associate professor in the Graduate School of Journalism at the University of British Columbia. His research interests include online journalism, social media and emerging genres of journalism.

*This research was supported by the Office of the Vice President for Research at the University of Minnesota and the Social Sciences and Humanities Research Council of Canada.*

science research (Lazer et al., 2009), while also raising troubling questions about consumer privacy, research ethics, and the overall quantification of social life (boyd & Crawford, 2012; Oboler, Welsh, & Cruz, 2012). For communication researchers in particular, the dramatic growth of social network sites has provided an ocean of data that reflect new media activities and artifacts: tweets, status updates, social recommendations, and more. With the Twitter platform especially, much of this information is freely available, making it a tempting target for researchers who can quickly “scrape” such data according to certain parameters. This, in turn, allows researchers to explore novel means of analyzing media content, as they use computational methods to assemble, filter, and interpret much of the collective Twitter conversation around a particular topic or event, such as the 2011 Arab Spring (Papacharissi & de Fatima Oliveira, 2012) or the 2010 federal election campaign in Australia (Burgess & Bruns, 2012).

This article focuses on the implications of the Big Data phenomenon for communication and media research. In particular, we examine the role of quantitative content analysis, one of the most familiar methods in mass communication (Berelson, 1952; Krippendorff, 2004; Riffe, Lacy, & Fico, 2005), within the context of Big Data. Scholars have long discussed the place of traditional, systematic content analysis in a digital information environment (McMillan, 2000; Weare & Lin, 2000), in some cases arguing that non-traditional variations are needed to cope with the unique nature of the Internet and its content (Herring, 2010; Karlsson, 2012). More recently, at the intersection of computer science and social science, communication scholars have incorporated computational techniques to study massive databases of media texts, especially the content flows of social media (e.g., Bruns & Burgess, 2012). While computerized content analysis has been around for decades (Riffe et al., 2005), the scope of emerging methods suggests that a re-examination of content analysis for communication research is necessary. On the one hand, traditional forms of manual content analysis were not designed to handle huge datasets of media texts (see Holsti, 1969); on the other hand, algorithmic analyses of content remain limited in their capacity to understand latent meanings or the subtleties of human language (Conway, 2006; Simon, 2001). Thus, this article asks: How can researchers best leverage the systematic rigor and contextual awareness of traditional content analysis while taking advantage of the efficiencies of computational methods and the novel cases offered by the Big Data phenomenon?

This article addresses that question by drawing on our own experience conducting a content analysis with a large cache of Twitter data, revealing the difficulties we encountered and the solutions we deployed—some algorithmic, some manual. We describe and reflect upon our research journey as we examined Andy Carvin’s use of Twitter as a “virtual newsroom” during the Arab Spring. Carvin, a social media strategist for NPR, emerged as a key broker of information on Twitter during the 2011 revolutions (Silverman, 2011). A study of his sourcing practices—i.e., whom he retweeted and how frequently he cited certain actor types—can begin to reveal the changing nature of journalistic sourcing in a networked age. The appeal for researchers is that the interactions between a reporter and source can be

tracked on Twitter, allowing for new research possibilities in revealing dynamics of influence and power. We argue that in many cases, scholars may see more fruitful results from a blended approach to content-analyzing Big Data—one that combines computational and manual methods throughout the process. Such an approach can retain the strengths of traditional content analysis while maximizing the accuracy, efficiency, and large-scale capacity of algorithms for examining Big Data.

## Literature Review

### Content Analysis in an Era of Big Data

The classical definition of content analysis is “a research technique for the objective, systematic, and quantitative description of the manifest content of communication” (Berelson, 1952, p. 18). Although several alternative definitions have been offered since, most scholars continue to agree that content analyses should be objective and systematic—that is, that each step must consistently follow explicitly formulated rules and procedures—although there is less agreement over the need for content analyses to be quantitative and to only analyze manifest content (Holsti, 1969; Titscher, Meyer, Wodak, & Vetter, 2000).

Content analyses generally conform to the following procedure: first, the research questions and/or hypotheses are formulated; second, the sample is selected; third, categories are defined for coding; fourth, coders are trained, the content is coded, and reliability is assessed; and fifth, the coded data are analyzed and interpreted (McMillan, 2000; Riffe et al., 2005). Although much of the scholarly literature is in agreement that the Web introduces a number of challenges to traditional conceptualizations of content analysis, there is less consensus over whether classical approaches are adequate for studying online content. McMillan (2000) and Weare and Lin (2000), for example, identify a number of challenges to applying content analysis to the Web, including difficulties in obtaining a representative sample because of the vastness of the Web; in defining the unit of analysis; and in ensuring that coders are presented with the same content for purposes of reliability. However, both ultimately conclude that minor adaptations to traditional approaches of content analysis, such as using lists to help generate sampling frames and using software to capture snapshots of Websites, are sufficient.<sup>1</sup>

The era of Big Data has been associated with a development that is simultaneously enticing and vexing for Internet researchers: massive information on and about human communication—a veritable “siren-song of abundant data” (Karpf, 2012, p. 10). For content analysts in particular, the development of digital media’s architecture has led to a fast-changing array of new *structural features* associated with communication, such as the hashtag on Twitter, as well as the development of *socio-cultural contexts* around those new features—both representing virgin terrain for content-analysis exploration.

This opportunity has attracted all manner of content analysis work. In one small but flourishing branch of this research alone, there are content analyses of Twitter use by journalists (Bruns, 2012b; Herrera & Requejo, 2012; Lasorsa, Lewis, & Holton, 2012), news organizations (Blasingame, 2011; Greer & Ferguson, 2011; Messner, Linke & Eford, 2012; Newman, 2011), foreign correspondents (Bruno, 2011; Cozma & Chen, 2012; Heinrich, 2012), nonprofit organizations (Waters & Jamal, 2011), and even the homeless (Koepller & Fleischmann, 2012). In such studies, researchers often content-analyze both the structural features of social media and their related socio-cultural contexts—as in Lasorsa and colleagues' examination of the unique retweet and reply functions of Twitter as a gauge of journalists' willingness to bring users into the gatekeeping process of news production.

Yet, this Internet development is also vexing because “the glittering promise of online data abundance too often proves to be fool's gold” (Karpf, 2012, p. 652). Researchers struggle to anticipate when and how to “trap” such data streams, given that most objects of study (e.g., tweets, Web traffic data, online news homepages) nearly all go un-archived (Karlsson & Strömbäck, 2010). Additionally, public data often pales in quality to proprietary data, which is rarely available to scholars. As boyd and Crawford (2012, p. 669) note, even while scholars are able to access vast numbers of tweets via Twitter's public Application Programming Interface (API), many researchers are not getting the “firehose” of the complete content stream, but merely a “gardenhose” of very limited numbers of public tweets—the randomness of which is entirely unknown, raising questions about the representativeness of such data to all tweets, let alone to all users on the service.

## Computational Methods as a Solution

If Big Data has thus provided communication scholars with a sea of digital information to content analyze, it also has yielded a series of digital enhancements to social science research: an emerging assemblage of tools and techniques for managing and making sense of all this data, often with no more than simple software on a standard computer (Manovich, 2012). More broadly, such computational methods can be bracketed within the phenomenon described at the outset of this article: a cultural, technological, and scholarly moment marked by two particular features—massive datasets containing digital traces of human behavior, and vastly lower barriers to organizing and analyzing such datasets (boyd & Crawford, 2012). Big Data simultaneously raises troubling questions about epistemology (e.g., how is knowledge constituted?) and ethics (e.g., how will user privacy be protected?), while also bringing with it an aura of technological destiny, heralding a new dawn in social science research. As such a notion has been described: “We no longer have to choose between data size and data depth” (see Manovich, 2012, p. 466).

This view helps to illustrate a fundamental tension for content analysis in an era of Big Data. In decades past, when content analysts faced the problem of too much data—too many news articles or television transcripts to code—they often resorted

to one of two tactics: They used random or stratified sampling methods to reduce data size, as in the constructed weeks approach described by Riffe et al. (2005), or they simply hired more student coders to do the work (Holsti, 1969). Computational methods, in theory, offer the potential for overcoming some of the sampling and coding limitations of traditional content analysis. First, with regard to sampling, this involves drawing on algorithmic measures for systematically gathering the entirety of the data of interest—such as the nearly 75,000 news articles published by a news organization in a calendar year, a task too unwieldy for any human to accomplish (Sjøvaag & Stavelin, 2012, p. 216). Thereafter, algorithmic techniques can be used to “slice” a vast corpus of data into smaller pieces for specialized analyses—as in the automatic timestamping of blog posts that can allow researchers to explore discussion topics evident during specific time periods (Bruns, Burgess, Highfield, Kirchhoff, & Nicolai, 2011, p. 280). Second, with regard to coding, this data can be examined using textual analysis and concept mapping tools that identify the most frequently used keywords and visualize their co-occurrence. In the case of Bruns et al. (2011), their computational assessment of blog posts shed new light on the networked nature of political discussion online.

However, when turning to computerized forms of content analysis, many scholars have found them to yield satisfactory results only for surface-level analyses, thus sacrificing more nuanced meanings present in the analyzed texts (Conway, 2006; Linderman, 2001; Nacos et al., 1991). Indeed, as Simon (2001, p. 87) notes, “The chief disadvantage is that the computer is simply unable to understand human language in all its richness, complexity, and subtlety as can a human coder.” This begs the question: what is the place of traditional forms of human-coded content analysis amid the allure of Big Data techniques?

## Blending Computational and Manual Methods

Another set of studies (Sjøvaag, Moe, & Stavelin, 2012; Sjøvaag & Stavelin, 2012) suggests that computational and manual approaches may be fruitfully combined within a single analysis, each type complementing the other. In their examination of public service news on the Web, Sjøvaag and colleagues blended computer-assisted data gathering and structuring with traditional quantitative content analysis to study a year’s worth of news content produced by the Norwegian Broadcasting Company (NRK). Because “freezing the flow” is imperative for studying online news (Karlsson & Strömbäck, 2010), Sjøvaag and her colleagues captured both a whole year’s worth of coverage—74,430 text-based news articles—and selected days of front-page images and information. The first dataset was computationally collected via Python scripts that downloaded the stories to a local server, then computationally analyzed through the automatic counting of Web-specific features such as hyperlinks, accommodation of reader comments, and videos. The second dataset, including roughly 2,000 articles, was quantitatively and manually coded, focusing on contextual features such as topics, news genre, thematic linking, and

sidebar content—none of which could be automatically recognized by software. “Computational power,” they argue, “ensures quality, precision and scale in registering platform-specific elements . . . while tried and tested practices of the news content analysis ensure assessment of thematic categorization” (Sjøvaag et al., 2012, p. 93).

The work of Sjøvaag and colleagues indicates that a promising way forward leverages computer-assisted methods for the coding of manifest content, while acknowledging that “human labor is still considered superior for the coding of latent content” (Sjøvaag & Stavelin, 2012, p. 219). For the literature on content analysis and its evolution, this approach suggests that the *structural features* of new media can, and should, be more fully subjected to algorithmic analysis, while the *socio-cultural contexts* built up around those features need the careful attention of manual methods. Moreover, while it’s true that computational techniques can be applied to the sampling and coding problems facing content analysis in a sea of too much information, we argue that, in many cases, scholars may be better served with a hybrid approach—one that *blends computational and manual methods throughout the content analysis process*.

## Case Study

Our case study built on the literature on news sourcing by applying it to the social media space, and focused on Andy Carvin’s notable role within that sphere of emerging journalistic practice online.

## Social Media and Journalistic Sourcing

The study of journalistic sourcing is key to an understanding of the media’s role in the construction of social reality. News sources not only influence much of the information obtained by journalists, but also critically provide a framework for interpreting it (Sigal, 1973). Journalists tend to cite individuals and organizations that hold positions of power in society, as they are considered authoritative and credible (Tuchman, 1978), leading to what Hall et al. describe as a “systematically structured over-accessing to the media of those in powerful and privileged institutional positions” (1978, p. 58). Because mainstream media generally marginalize voices considered deviant, such groups often are “literally rendered speechless” (Cottle, 2003, p. 6). But as Manning (2001) found in the case of groups such as Greenpeace and Friends of the Earth, marginalized sources may later be considered authoritative and move up the hierarchy of credibility.

Digital technologies such as social media have attracted the interest of scholars as alternative platforms of public communication. As a channel for news dissemination, Twitter has become a convenient tool for journalists to interact with possible sources and gather information from a vast pool of people without leaving the office or even

picking up the phone (Hermida, 2010, 2011). Our approach builds on the notion of Twitter as a beat advanced by Broersma and Graham (2012, p. 405), defined as “a virtual network of social relations of which the journalist is a part with the purpose of gathering news and information on specific topics.” But we go further to conceive of Twitter as both newswire and newsroom. As a newswire, Twitter provides a constantly updated public source of raw material in near real-time. As a newsroom, it offers a collaborative public space for the processing of information through the application of journalistic techniques such as evaluating reports, verifying facts and highlighting the most relevant information.

In contrast to the gatekeeping processes within mainstream media newsrooms, the selection of relevant content in social media spaces occurs after, not before, publication—and it is determined by distributed users, not centralized editors, through mechanisms such as the retweet on Twitter. Papacharissi and de Fatima Oliveira (2012) noted the emergence of alternative voices as prominent sources during the Egyptian uprising, crowdsourced via the retweet function. Poell and Borra (2012) suggest that the specific attributes of Twitter hold the most promise for crowdsourced alternative reporting. These findings point to the need for employing “natively digital” (Rogers, 2009) approaches to evaluating social media texts for content analysis—that is, evaluating the relevance of these new media objects according to markers of aggregate relevance (most-clicked, most-retweeted, etc.), rather than more traditional attempts such as word-frequency analyses (Poell & Borra, 2011, p. 700).

The technical architecture of Twitter thus allows researchers to investigate both the newswire and newsroom aspects of this space. Sources cited are captured by the retweet mechanism, when a journalist cites a message from an individual or organization. While the sourcing process is visible, it is difficult to grasp given the volume, speed, and ephemeral nature of the data. The interactions between a journalist and a source are captured by the @mentions mechanisms, uncovering how a journalist engages with sources to gain information, background, and context. The @mentions interactions amount to a public newsroom where exchanges can be tracked, recorded, and analyzed—in contrast to traditional interactions with sources that take place away from public view.

## Andy Carvin's Arab Spring Coverage as a Case Study

By analyzing the Twitter feed of NPR's Carvin during critical periods of the uprisings in Tunisia and Egypt, we sought to contribute to a growing body of work that investigates how social media, and more specifically Twitter, can offer a platform for the co-construction of news by journalists and the public, expanding the range and type of actors in the news (Lotan et al., 2011; Papacharissi & de Fatima Oliveira, 2012).

We chose to focus on Carvin's work due to his role as a vital broker of information on Twitter during the Arab Spring (Silverman, 2011). His work provided a case study to discover if a new journalistic style was at play in social spaces, where reporters

rely on and interact with a potentially broader range of sources. We sought to identify two key variables: firstly, the type of interaction, based on whether the interaction was a retweet or @mention; secondly, the type of source being interacted with, adapting the actor type classifications from Lotan et al. (2011).

Our findings indicated that non-elite sources had a greater influence over the content flowing through his Twitter stream than journalists or other elite sources (Hermida, Lewis, & Zamith, forthcoming). While non-elite actors barely made up a quarter of his sources overall, they accounted for nearly half of all the messages in the sample. Our analysis showed that Carvin gave more prominence in his retweets to alternative voices, compared to journalists or other elite sources. By contrast, his interactions through the use of @mentions suggest he was engaging in an ongoing discussion about the emerging genres of digital journalism, rather than a form of newsgathering.

## Method

In order to analyze a substantial amount of data while still remaining sensitive to contextual nuance, we opted to blend computational and manual methods. In particular, we created a Python script to categorize a large dataset; used spreadsheet and statistical software to organize the data and identify the objects of our analysis; converted dynamic Web pages into static objects with open-source software; and developed a Web-based electronic coding interface to facilitate the work of human coders and reduce error. The sections below describe this work of obtaining, parsing, interpreting, and, ultimately, content-analyzing the data.

### Obtaining Data

Twitter has become a popular medium for researchers in large part because the Twitter API makes it relatively easy to scrape, or download, massive numbers of tweets—literally hundreds of thousands of messages (e.g., Lotan et al., 2011). While recent changes to the API have limited researchers' access to public tweets, it remains possible to gather tweets organized by a keyword or hashtag (Bruns, 2012a; Bruns & Burgess, 2012). In our case, however, we faced the particular challenge of collecting all the tweets generated by a single user (@acarvin; see <http://twitter.com/acarvin>). Custom-written Python scripts can begin to accomplish this task, but, in our experience, they often "time out" after gathering only a small number of a user's public tweets. The first author thus tried to use a combination of RSS feeds and the TwapperKeeper service (now discontinued), but these also rendered only selections of Carvin's entire set of tweets. Thereafter, the first author learned that Carvin had obtained from Twitter directly a file containing all of his more than 60,000 tweets from December 2010 to September 2011; Carvin agreed to share the dataset with us for this analysis.

Figure 1  
Screenshot of the Data File Provided by NPR's Andy Carvin

```
535031 ***** latitude38.9977906784geo_precision153name6417871953fa5e86long.
535032 geo: latitude38.9977906784geo_precision153name6417871953fa5e86longitude-77.0365996008enti:
535033 annotations: 778057 oauth:129032 false false false 298240940 Fri Jun 03 12:48:17 +0000
535034 user_id: 778057 oauth:129032 false false false 298240940 Fri Jun 03 12:48:17 +0000
535035 media: oauth:129032 false false false 298240940 Fri Jun 03 12:48:17 +0000
535036 created_via: oauth:129032 false false false 298240940 Fri Jun 03 12:48:17 +0000
535037 has_takedown: false false false 298240940 Fri Jun 03 12:48:17 +0000 2011 "not dead then, eh? "@Reuters: FLASH:
535038 nsfw_user: false false 298240940 Fri Jun 03 12:48:17 +0000 2011 "not dead then, eh? "@Reuters: FLASH:
535039 nsfw_admin: false 298240940 Fri Jun 03 12:48:17 +0000 2011 "not dead then, eh? "@Reuters: FLASH:
535040 id: 298240940 Fri Jun 03 12:48:17 +0000 2011 "not dead then, eh? "@Reuters: FLASH:
535041 created_at: Fri Jun 03 12:48:17 +0000 2011 "not dead then, eh? "@Reuters: FLASH:
535042 contributor_id: "not dead then, eh? "@Reuters: FLASH: Yemeni president Saleh alive, will hold news
535043 "text: not dead then, eh? "@Reuters: FLASH: Yemeni president Saleh alive, will hold news
535044 #VALUE! source_user_id19360470source_status_id76630433693040641parent_status_id76630433693040641pa
535045 ***** source_user_id19360470source_status_id76630433693040641parent_status_id76630433693040641pa
535046 annotations: source_user_id19360470source_status_id76630433693040641parent_status_id76630433693040641pa
535047 share: source_user_id19360470source_status_id76630433693040641parent_status_id76630433693040641pa
535048 user_id: 778057 oauth:129032 false false false 298231400 Fri Jun 03 12:49:22 +0000
535049 media: oauth:129032 false false false 298231400 Fri Jun 03 12:49:22 +0000
535050 created_via: oauth:129032 false false false 298231400 Fri Jun 03 12:49:22 +0000
535051 has_takedown: false false false 298231400 Fri Jun 03 12:49:22 +0000 2011 "not dead then, eh? "@Reuters: FLASH:
535052 nsfw_user: false false 298231400 Fri Jun 03 12:49:22 +0000 2011 "RT @batt
```

While obtaining data straight from the source may not be an option for most researchers,<sup>2</sup> it is important to note that, as we compared what the first author initially scraped and what Carvin provided us, we found significant discrepancies. The researcher-scraped sample was missing “chunks” of tweets during some time periods, and it included less metadata or other information about the tweets. These gaps reinforce the limitations of relying on algorithmic methods for scraping tweets when the comprehensiveness of such data collection is in question (boyd & Crawford, 2012). This is an important consideration because using an incomplete dataset will have a negative impact on the content validity and perhaps other forms of internal validity of a study. Researchers should thus strive to acquire the most comprehensive data available, which will often come directly from the source. In our example, the researcher-scraped sample did not suffer from random error; rather, the missing information was in “chunks,” potentially leading to disproportionate effects on certain source types and sourcing practices, and consequently on the validity of our study.

Parsing Data

The file provided by Carvin was a 199-megabyte, tab-delimited text file. Each entry contained 11 unique identifiers from Twitter’s database: *Annotations*, *User ID*, *Media*, *Created Via*, *Has Takedown*, *NSFW User*, *NSFW Admin*, *ID*, *Date*, *Contributor ID*, *Text*, and *Filter*. However, some entries included specific identifiers that others did not, such as the geographic location of the user at the time the tweet was submitted. Furthermore, as shown in Figure 1, this file offered limited adherence to the conventions of CSV-like tabular, plain-text data files, thus preventing us from

easily importing them into a spreadsheet or statistics program for analysis. Lastly, there appeared to be no identifiers for variables that were key to our analysis, such as whether a tweet was a retweet or not.

In light of these difficulties and shortcomings, as well as the large size of the data set, we had to come up with a computer-assisted solution that would extract the most important data from the text file and organize it in a manner conducive to a content analysis. A computer-assisted solution was appropriate here because computer programs are efficient at identifying and categorizing information based on structural features and systematic patterns, and are able to follow unambiguous instructions with exact precision (Leetaru, 2012). Indeed, while a graduate assistant (as Holsti [1969] might suggest) may have been able to go through this file and systematically copy and paste the relevant data into a separate, cleaner file, it likely would have been prohibitively time-consuming and the result subject to mistakes. We thus composed a Python script to help accomplish this task. We selected Python because it is a simple, yet powerful, high-level, general-purpose programming language that is well-suited to developing scripts for parsing through large bodies of mostly-uniform data.

The first thing we did was identify the components that were crucial to our analysis. These were: (1) the date of the tweet, (2) the body text of the tweet, (3) the usernames of the sources mentioned in each tweet, (4) whether a tweet was a retweet or not, and, (5) if it was a retweet, the username of the source being retweeted. Having identified these components, we then looked for patterns within the data file that would allow us to extract and organize the data.

First, we had to isolate each tweet into a separate unit of analysis. In reviewing the data file, we noticed that every unique entry began with a sequence of 20 asterisks, and thus programmed our script to use that as a delimiter. The date of the tweet was provided in a standard fashion in all entries in the file: *Weekday Month Day Hour:Minute:Second +UTC Year*. We had our script identify the date of each entry by looking for the first instance of the text "created\_at:" within each entry and reorganize the subsequent information into a form that was easier to read by other programs: *YYYY-MM-DD*. This was necessary in order to enable us to later filter through the tweets with ease and accuracy, and quickly select only the tweets that appeared on the dates in our sampling frame. We similarly identified the body of the tweet by extracting all of the characters present between text "text:" and the first tab following it.

In order to obtain a list of all of the sources appearing in a tweet, we programmed our script to extract all characters appearing between the "@" character and a blank space each time the "@" character appeared in the body. This process was necessary because individual tweets often included multiple sources, and our script thus created a comma-delimited list of all sources appearing in the tweet in order to enable us to later accurately analyze the prominence of individual sources and to segment a specific manner of online interaction: engagement. Retweets were identified by the presence of the text "RT @" in the body of the tweet, and were automatically coded by our program as 1 (retweet) or 0 (not a retweet). This

Figure 2  
Screenshot of the Excel Spreadsheet with All Tweets

	A	B	C	D	E	F	I
1	ID	Date	Time	Body	RT?	RTed Source	Sources
131	1055	1/13/2011	19:18:02	@kategardiner, @robinsloan: police sniper	0	None	@kategardiner, @robinsloan
132	1056	1/13/2011	19:18:46	@robinsloan: if you look near the bottom o	0	None	@robinsloan
133	1057	1/13/2011	19:19:43	@KateGardiner: people I know are tweetin	0	None	@KateGardiner
134	1058	1/13/2011	19:20:13	RT @Dima_Khatib: Ben Ali: I've instructed n	1	Dima_Khatib	@Dima_Khatib
135	1059	1/13/2011	19:21:20	What's your source? Need to confirm. RT @	1	karim2k	@karim2k
136	1060	1/13/2011	19:23:12	@jilliancyork: have you seen any confirmat	0	None	@jilliancyork
137	1061	1/13/2011	19:43:42	A curfew imposed in Tunis offers the youth	0	None	None
138	1062	1/13/2011	19:49:13	RT @simst: For the first time in 23 years,	1	simst	@simst
139	1063	1/13/2011	19:56:58	@jilliancyork: Thanks... Seemed the tweets	0	None	@jilliancyork

was necessary in order to segment another specific manner of online interaction between Carvin and his sources, with retweets serving as a form of broadcasting. The distinction between broadcasting and engagement was central to one of our research questions; without this segmentation, such line of inquiry would not have been possible. We also identified the username being retweeted by programming our script to extract all characters appearing between the “RT @” text and a blank space. However, unlike the aforementioned procedure for identifying all sources, our program only extracted the text following the first match. This was done because the message being retweeted would sometimes be a retweet itself. However, our study was concerned with the individuals that Carvin sought to retweet, and not the chain of retweets, thus requiring us to make this distinction.

For each of the 60,114 units (tweets) extracted from the original data file, our script created a new line in a CSV file and output all of our variables onto that line in a uniform fashion. This provided us with a single file containing all of the data, which was later imported into Microsoft Excel.<sup>3</sup> As shown in Figure 2, this new arrangement allowed us to easily filter through the data and presented it in a form that was easier for us to interpret, thereby facilitating our ability to look at the “big picture” and reducing the likelihood for researcher error, which might have threatened the validity of our study.

Interpreting and Filtering Data

Once the data had been reorganized and verified through spot-checking, we began to filter through it to obtain the sample that was most relevant to our study. First, we filtered the dates to only show tweets appearing between January 12 and January 19 (Tunisian revolution), and copied those entries to a separate spreadsheet. We repeated this process for the dates of January 24 to February 13 (Egyptian revolution). For each of these two files, we segmented between the retweeted messages (broadcast) and non-retweeted messages (engagement) and, using the statistical package SPSS, ran descriptive analyses to identify all of the unique sources and the proportion of the overall sources that they represented within the respective

segment. Again, while this process may have been performed by humans, it would have likely taken them several hours and the result may have been marred by mistakes. In contrast, our computer-assisted approach was not only more reliable, but the task was completed in just minutes. This information was saved into a separate CSV file, which we later cross-referenced just prior to commencing the statistical analysis.

We thus ended up with four sets of information: Tunisia Broadcast, Tunisia Engagement, Egypt Broadcast, and Egypt Engagement. To create a comparable and sufficiently large, yet manageable, sample, the researchers opted to code all profiles that accounted for 0.09% or more of the retweeted sources or 0.25% or more of the non-retweeted sources. The decision to sample only the most prominent sources reflects the limitations of human coding, which we discuss in the following subsection. While a fully automated solution may have yielded insight into the “long tail,” which we admittedly missed, it would have likely compromised the validity of our study as a result of its limited sensitivity to context. Because some sources were present in multiple conditions, we once again used SPSS to identify all of the unique sources appearing in the four conditions. Our final sample was comprised of 330 unique sources.

## Content Analysis

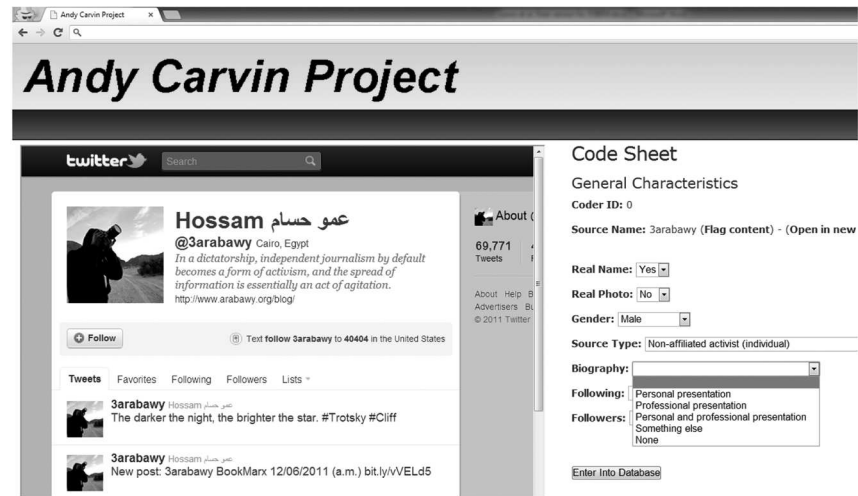
While the works of scholars like Bruns and colleagues have illustrated different ways in which computational methods may be leveraged to sort, categorize, and show links between texts and their content based on structural features and surface-level analyses, their use remains somewhat limited when more latent features are of concern. Indeed, although the fields of natural language processing, computational linguistics, and text analytics continue to mature, they arguably remain unable to match the ability of humans to take context into account and make fine distinctions (Conway, 2006; Simon, 2001). In our case, we sought to classify individual sources into singular classes, a rather challenging task given the difficulty of making such judgments when relying primarily on the brief, self-disclosed biographical blurb present in a user’s profile page (Lotan et al., 2011). Because of this ambiguity, a computational approach likely would have yielded unsatisfactory results; indeed, our own coders sometimes required additional context, which they obtained by reviewing outside resources, such as personal Web pages linked to from the biographical blurb on the source’s profile. In light of this, we opted to conduct a human-driven content analysis.

Our first task, therefore, was to obtain the Twitter profile page of each source in order to provide coders with the necessary information and context to make decisions. As noted by Karpf (2012), online texts are malleable, and oftentimes constantly changing, thereby presenting a challenge to intercoder agreement, since there can be no guarantee that multiple coders would review the same material. This is especially true of Twitter profile pages, which are constantly updated by

both the user (in the form of new tweets and the alteration of profile information) and the service (fluctuations in the numbers of followers and followees). In light of this consideration, we created a simple script to download the Twitter profile of each of the unique sources and store them as separate files. To accomplish this, we generated a list of all of the unique usernames and then used a popular, open-source program called Wget to systematically download the profile pages of each user ([http://www.twitter.com/\[username\]](http://www.twitter.com/[username])) and name the resulting file after the username (e.g., ghonim.html). This procedure allowed us to archive our own static copies of the profile pages, thereby ensuring that coders would be examining the exact same content.

Once we had obtained the profile pages for all of our sources, we developed a Web-based electronic coding interface using the PHP scripting language and a MySQL database. In this system, the source's archived Twitter profile page appeared adjacent to the electronic code sheet. Individual sources were pre-assigned to coders to ensure equal distribution by creating a spreadsheet in Microsoft Excel that linked sources with specific coders, exporting that list as a CSV file, and generating a MySQL table with that information by using the import feature of phpMyAdmin, a Web-based administration interface for MySQL. Thus, once a coder signed on, the system located the next source on their assignment list and automatically called up the archived profile page and presented it to the coder alongside the code sheet. As shown in Figure 3, the code sheet was comprised of interactive, human-readable variables and categories. Coders could also click on the variable names in order to be presented with the codebook definition for that variable. When the coder

Figure 3  
Screenshot of the Web Interface for Content Analysis



finished coding a source, the information was automatically entered into a separate MySQL table. Using phpMyAdmin, data from that table was then exported using phpMyAdmin and opened in SPSS for analysis.

This electronic setup was useful for five main reasons. First, it helped reduce coder error by having options and categories presented as labels, with the system automatically converting selections into numerical values after each submission. Second, it eliminated data entry error by automatically transferring entries to a relational database, removing the need for human intervention. Third, it enabled the researchers to quickly conduct multiple rounds of intercoder reliability by simply copying and pasting new lists of sources and refreshing the database, and subsequently transition into the full coding with ease. Fourth, it enabled coders to flag units, allowing them to either later return to a specific unit or to discuss their concerns with one of the researchers. Lastly, it was both convenient for the coders, allowing them to code at their convenience and easily transition between the material being coded and the code sheet, and facilitated the ability of the researcher to monitor and review the work of the coders. Ultimately, this setup was used for multiple rounds of intercoder reliability testing and in the main coding, with our coders expressing satisfaction with the approach.

## Discussion

In this article, we have articulated the challenges and opportunities of Big Data for systematic, quantitative content analysis in media and communication research. As the shift from mass communication to networked communication gives rise to ever-larger bodies of public communication online, the resulting datasets—massive in size, daunting in scope—have strained the capacity of traditional content analysis, encouraging media content analysts to find innovative solutions to the persistent challenges of selecting and coding texts. Computational methods offer new pathways for addressing such problems, and yet there are limitations to what they alone can accomplish. We thus argue that there are many cases where scholars are best served with a hybrid approach that blends computational and manual methods throughout the content analysis process. Our case study of Andy Carvin illustrates this hybrid approach in action. Computational methods and tools were used to objectively, systematically, and accurately filter the sample, while facilitating the work of human coders by removing several inefficiencies and potential for data-entry error that are typical in manual coding. Through it all, computational means were enlisted to enhance, rather than supplant, the work of human coders, enabling them to tackle a larger body of data while remaining sensitive to contextual nuance. In this way, hybrid combinations of computational and manual approaches can preserve the strengths of traditional content analysis, with its systematic rigor and contextual awareness, while maximizing the large-scale capacity of Big Data and the efficiencies of computational methods.

There are additional ways in which computational methods can be leveraged to improve the work of human coders. Researchers may use algorithmic methods to provide human coders with additional cues, thereby helping to promote intercoder reliability. For example, in a traditional content analysis, researchers may give coders a set of keywords as indicators for a given category; a hybrid approach could automatically perform a search for such keywords within a given text and subsequently suggest the respective category to a human coder, who would then either confirm or revise the suggested assessment. Such an approach could be even more powerful when combined with natural language processing tools, which are better suited to handle complex instructions and are more sensitive to context.

However, in the allure of computational methods, researchers must not lose sight of the unique role of humans in the content analysis process. This is particularly true of their ability to bring contextual sensitivity to the content, its manifest and latent characteristics, and its place within the larger media ecology. This contextual awareness is doubly important in studies of Big Data, where the sheer volume might encourage researchers to look for aggregate patterns—while missing the subtleties embedded in the production of that digital content. As boyd and Crawford (2012) caution, “Context is hard to interpret at scale and even harder to maintain when data are reduced to fit into a model. Managing context in light of Big Data will be an ongoing challenge” (p. 671).

The peculiar and ironic challenge for content analysis in this sphere is that Big Data often is composed of “small data”—many thousands of tiny, discrete traces of digital communication, such as 140-character tweets. These micro messages typically come from many different users who are creating content for different purposes and for different intended audiences (Marwick & boyd, 2011). This creates problems on two levels. Firstly, whereas content analysts traditionally have had longer texts, like newspaper articles, through which to assess framing and tone, the brevity of tweets may require additional work to determine meaning. Secondly, because the many small data within Big Data may represent a wide variety of user intentions, human coders have an important role in reclaiming context to the extent possible—for example, understanding the rhythms of Twitter or the culture around a particular hashtag well enough to settle on a valid and reliable set of measures for analysis. Indeed, the media logic of Twitter demonstrates the challenge this poses for content analysis. Because tweet streams are multi-faceted and fragmented, the significance of the data lies less in a single piece that may be of limited validity or importance on its own and more in the aggregate effect of the messages (Hermida, 2010). The collective nature of Twitter means that a single tweet from an individual may be relevant at a specific moment for a specific reason, whereas the remainder of the stream may be irrelevant.

The media logic of Twitter also points to another challenge to context. In our case, we sought to classify the actor types. Had we conducted this study with a newspaper sample, the reporter would have defined the actor type (e.g., labeling a source as a police officer or activist). On Twitter, users are able to self-define their identity, rather than have a classification imposed on them. For our study,

we relied on a user's Twitter bio as the primary indicator of identity. Yet, it was hard to classify a considerable number of users who spanned some of the categories. Lotan et al. (2011) experienced the same difficulty, suggesting that "many influential Twitter users do not easily fit into traditional categorization schemes" (p. 1397). We suggest that the issue reflects the differences between closed, professionally produced media where identity is assigned to a source, and open, collaborative media platforms where identity is selected by the user (cf., Lewis, 2012).

In the end, there are trade-offs when applying manual content analysis to Big Data—either a reduction in sample, or an increase in human coders—just as there has always been with traditional content analysis of mass communication (Holsti, 1969). In our case, faced with a mountain of data, we sought to reduce it to a level that was feasible for human coding. But, in contrast to traditional approaches, we leveraged computational methods to magnify what a small set of coders could accomplish, enabling them to more efficiently and accurately identify the context of Twitter actors. This, we argue, is the essence of reinventing traditional content analysis for the era of Big Data. As scholars increasingly take up such datasets, they may be better served by interweaving computational and manual approaches throughout the overall process in order to maximize the precision of algorithms and the context-sensitive evaluations of human coders.

## Notes

<sup>1</sup>It is worth noting that this view is challenged by scholars like Herring (2010) and Karlsson (2012), who share a concern that the structural features of new media (such as hyperlinks) and the media content created through them (such as blog comments) are simply too "new" to be addressed by "old" methods of content analysis alone.

<sup>2</sup>As of late December 2012, Twitter introduced a feature allowing users to download their entire Twitter archive. Thus, researchers studying an individuals' (or set of individuals') use of Twitter, as we did, may request from the subjects of their study a copy of their Twitter archive. Because this feature was not available to us at the time of writing, we have only briefly experimented with it. However, according to the documentation accompanying the download (as of January 2013), "the JSON export contains a full representation of your Tweets as returned by v1.1 of the Twitter API." As such, it would be a rich and comprehensive data set, like the one provided to us by Carvin, available in a standardized format that is easily parsed.

<sup>3</sup>Although we used Microsoft Excel and SPSS because of the researchers' collective familiarity with those programs, our approach may be easily adopted using open-source solutions, such as LibreOffice and R, respectively.

## References

- Berelson, B. (1952). *Content analysis in communication research*. New York, NY: Free Press.
- Blasingame, D. (2011). Gatejumping: Twitter, TV news and the delivery of breaking news. *#ISOJ Journal: The Official Research Journal of the International Symposium on Online Journalism*, 1.

- boyd, d. & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679. doi: 10.1080/1369118X.2012.678878
- Broersma, M., & Graham, T. (2012). Social media as beat: Tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice*, 6(3), 403–419. doi: 10.1080/17512786.2012.663626
- Bruno, N. (2011). Tweet first, verify later: How real-time information is changing the coverage of worldwide crisis events. Reuters Institute for the Study of Journalism. Retrieved from: [http://reutersinstitute.politics.ox.ac.uk/fileadmin/documents/publications/fellows\\_papers/2010-2011/TWEET\\_FIRST\\_VERIFY\\_LATER.pdf](http://reutersinstitute.politics.ox.ac.uk/fileadmin/documents/publications/fellows_papers/2010-2011/TWEET_FIRST_VERIFY_LATER.pdf)
- Bruns, A. (2012a). How long is a tweet? Mapping dynamic conversation networks on Twitter using Gawk and Gephi. *Information, Communication & Society*, 15(9), 1323–1351. doi: 10.1080/1369118X.2011.635214
- Bruns, A. (2012b). Journalists and Twitter: How Australian news organisations adapt to a new medium. *Media International Australia Incorporating Culture and Policy*, 144, 97–107.
- Bruns, A., & Burgess, J. (2012). Researching news discussion on Twitter: New methodologies. *Journalism Studies*, 13(5–6), 801–814. doi: 10.1080/1461670X.2012.664428
- Bruns, A., Burgess, J., Highfield, T., Kirchhoff, L., & Nicolai, T. (2011). Mapping the Australian Networked Public Sphere. *Social Science Computer Review*, 29(3), 277–287. doi: 10.1177/0894439310382507
- Burgess, J., & Bruns, A. (2012). (Not) the Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice*, 6(3), 384–402. doi: 10.1080/17512786.2012.663610
- Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, 83(1), 186–200. doi: 10.1177/107769900608300112
- Cottle, S. (2003). *News, public relations and power*. London, UK: Sage
- Cozma, R., & Chen, K.-J. (2012). What's in a tweet? Foreign correspondents' use of social media. *Journalism Practice*, 7(1), 33–46. Advance online publication. doi: 10.1080/17512786.2012.683340
- Greer, C. F., & Ferguson, D. A. (2011). Using Twitter for promotion and branding: A content analysis of local television Twitter sites. *Journal of Broadcasting & Electronic Media*, 55(2), 198–214. doi: 10.1080/08838151.2011.570824
- Hall, S., Critcher, C., Jefferson, T., Clarke, J. N., & Roberts, B. (1978). *Policing the crisis: Mugging, the state, and law and order*. London, UK: Macmillan.
- Heinrich, A. (2012). Foreign reporting in the sphere of network journalism. *Journalism Practice*, 6(5–6), 766–775. doi: 10.1080/17512786.2012.667280
- Hermida, A. (2010). Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3), 297–308. doi: 10.1080/17512781003640703
- Hermida, A. (2011). Tweet the news: Social media streams and the practice of journalism. In S. Allan (Ed.), *The Routledge companion to news and journalism* (pp. 671–82). New York, NY: Routledge.
- Hermida, A., Lewis, S. C. & Zamith, R. (forthcoming). Sourcing the Arab Spring: A case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian Revolutions. *Journal of Computer-Mediated Communication*.
- Herrera, S. & Requejo, J. L. (2012). 10 good practices for news organizations using Twitter. *Journal of Applied Journalism & Media Studies*, 1(1), 79–95.
- Herring, S. C. (2010). Web content analysis: Expanding the paradigm. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of Internet research* (pp. 233–249). The Netherlands: Springer.
- Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. Reading, MA: Addison-Wesley Publishing Company.
- Karlsson, M. (2012). Charting the liquidity of online news: Moving towards a method for content analysis of online news. *International Communication Gazette*, 74(4), 385–402. doi: 10.1177/1748048512439823

- Karlsson, M., & Strömbäck, J. (2010). Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news. *Journalism Studies*, 11(1), 2–19. doi: 10.1080/14616700903119784
- Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society*, 15(5), 639–661. doi: 10.1080/1369118X.2012.665468
- Koepfler, J. A., & Fleischmann, K. R. (2012). Studying the values of hard-to-reach populations: Content analysis of tweets by the 21st century homeless. *Proceedings of the 2012 iConference, iConference '12* (pp. 48–55). New York, NY: ACM. doi: 10.1145/2132176.2132183
- Krippendorff, K. (2004). *Content Analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications.
- Lasorsa, D. L., Lewis, S. C., & Holton, A. E. (2012). Normalizing Twitter: Journalism practice in an emerging communication space. *Journalism Studies*, 13(1), 19–36. doi: 10.1080/1461670X.2011.571825
- Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Gutmann, M. (2009). Life in the network: The coming age of computational social science. *Science*, 323(5915), 721–723.
- Leetaru, K. H. (2012). *Data mining methods for the content analyst: An introduction to the computational analysis of content*. New York, NY: Routledge.
- Lewis, S. C. (2012). The tension between professional control and open participation: Journalism and its boundaries. *Information, Communication & Society*, 15(6), 836–866. doi: 10.1080/1369118X.2012.674150
- Linderman, A. (2001). Computer content analysis and manual coding techniques: A comparative analysis. In M. D. West (Ed.), *Theory, Method, and Practice in Computer Content Analysis*. Westport, CT: Ablex Pub. Corp.
- Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & boyd, D. (2011). The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication*, 5, 1375–1405.
- Manning, P. (2001). *News and news sources: A critical introduction*. London: Sage.
- Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–475). Minneapolis, MN: University of Minnesota Press.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Retrieved from [http://www.mckinsey.com/insights/mgi/research/technology\\_and\\_innovation/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation)
- Marwick, A. E., & boyd, D. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1), 114–133. doi: 10.1177/1461444810365313
- Messner, M., Linke, M., & Eford, A. (2012). Shoveling tweets: An analysis of the microblogging engagement of traditional news organizations. *#ISOJ Journal: The Official Research Journal of the International Symposium on Online Journalism*, 2(1).
- McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly*, 77(1), 80–98. doi: 10.1177/107769900007700107
- Nacos, B. L., Shapiro, R. Y., Young, J. T., Fan, D. P., Kjellstrand, T., & McCaa, C. (1991). Content analysis of news reports: Comparing human coding and a computer-assisted method. *Communication*, 12, 111–128.
- Newman, N. (2011). Mainstream media and the distribution of news in the age of social discovery. Reuters Institute for the Study of Journalism. Retrieved from: [http://reutersinstitute.politics.ox.ac.uk/fileadmin/documents/Publications/Working\\_Papers/Mainstream\\_media\\_and\\_the\\_distribution\\_of\\_news\\_.pdf](http://reutersinstitute.politics.ox.ac.uk/fileadmin/documents/Publications/Working_Papers/Mainstream_media_and_the_distribution_of_news_.pdf)
- Oboler, A., Welsh, K., & Cruz, L. (2012). The danger of big data: Social media as computational social science. *First Monday*, 17(7–2). Retrieved from <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3993/3269>

- Papacharissi, Z., & de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication*, 62(2), 266–282. doi: 10.1111/j.1460-2466.2012.01630.x
- Poell, T., & Borra, E. (2012). Twitter, YouTube, and Flickr as platforms of alternative journalism: The social media account of the 2010 Toronto G20 Protests. *Journalism*, 13(6), 695–713. doi: 10.1177/1464884911431533
- Riffe, D., Lacy, S. R., & Fico, F. G. (2005). *Analyzing media messages: Using quantitative content analysis in research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rogers, R. (2009). *The End of the virtual — Digital methods*. Amsterdam: Amsterdam University Press.
- Sigal, L. V. (1973). *Reporters and officials: The organization and politics of newsmaking*. Lexington, MA: D. C. Heath.
- Silverman, C. (2011, April 8). Is this the world's best Twitter account? *Columbia Journalism Review*. Retrieved from [http://www.cjr.org/behind\\_the\\_news/is\\_this\\_the\\_worlds\\_best\\_twitter\\_account.php](http://www.cjr.org/behind_the_news/is_this_the_worlds_best_twitter_account.php)
- Simon, A. F. (2001). A unified method for analyzing media framing. In R. P. Hart & D. R. Shaw (Eds.), *Communication in U.S. elections: New agendas* (pp. 75–89). Lanham, MD: Rowman and Littlefield.
- Sjøvaag, H., Moe, H., & Stavelin, E. (2012). Public service news on the Web: A large-scale content analysis of the Norwegian Broadcasting Corporation's online news. *Journalism Studies*, 13(1), 90–106. doi: 10.1080/1461670X.2011.578940
- Sjøvaag, H., & Stavelin, E. (2012). Web media and the quantitative content analysis: Methodological challenges in measuring online news content. *Convergence: The International Journal of Research into New Media Technologies*, 18(2), 215–229. doi: 10.1177/1354856511429641
- Titscher, S., Meyer, M., Wodak, R., & Vetter, E. (2000). *Methods of text and discourse analysis*. London, UK: Sage Publications.
- Tuchman, G. (1978). *Making News: A study in the construction of reality*. New York, NY: Free Press.
- Waters, R. D., & Jamal, J. Y. (2011). Tweet, tweet, tweet: A content analysis of nonprofit organizations' Twitter updates. *Public Relations Review*, 37(3), 321–324. doi: 10.1016/j.pubrev.2011.03.002
- Weare, C., & Lin, W.-Y. (2000). Content analysis of the World Wide Web: Opportunities and challenges. *Social Science Computer Review*, 18(3), 272–292. doi: 10.1177/089443930001800304